MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 1, 2017/2018 SESSION

TDS 3301 – DATA MINING

(All Sections / Groups)

14 OCTOBER 2017
2:30 p.m – 4:30 p.m
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This Question paper consists of 4 printed pages including cover page with 4 questions only.
2. Attempt **ALL** questions. All questions carry equal marks and the distribution of marks for each question is given.
3. Please write all your answers in the Answer Booklet provided

Question 1

A survey was conducted to study the percentage of salary an adult spend on entertainment. The **sample** of 10 randomly selected adults are tabulated as follows.

Age	23	38	35	32	46
%Salary	16.1	17.3	27.2	19.8	22.7
Age	17	41	27	34	26
%Salary	13.6	20.2	20.7	20.5	15.6

Present your final answers in two decimal places.

a) Calculate the *mean* and *standard deviation* for age and % salary. (2 marks)
b) Normalize the age and % salary using *z-score normalization.* Present your answer in a tabulated form. (2 marks)
c) Use your answer in (b) to calculate the *correlation coefficient* between age and % salary. Show your steps and justify whether they are *positively* or *negatively correlated.* (4+1 marks)
d) Give a reason why *n-1* for calculating variance of a sample? (1 mark)

Question 2

a) *Subjective measure* is sometime used to determine the quality of association rules. Explain the measure in brief. (1 marks)
b) Objective measure, on the other hand, uses support and confidence to determine the quality of association rules. What are *support* and *confidence*? (2 marks)
c) What are *exact rules* and *strong rules*? (2 marks)
d) Explain *frequent item set* and *Apriori property* of Apriori algorithm. (2 marks)
e) A transaction database is given as follow. Let minimum support = 60% and minimum confidence = 80%. Generate *frequent item sets* using Apriori. Show your steps. (3 marks)

Transaction	Items
T1	{a,b,c,d,e,f}
T2	{g,b,c,d,e,f}
T3	{a,h,d,e}
T4	{a,i,j,d,e }
T5	{j,b,f,d,k,e}

Continued...

Question 3

a) What is the *simplified assumption* made for attributes used in the Naïve Bayes classifier (NBC)? (1 mark)
b) What is the *advantage* and *disadvantage* of making assumption as in (a)? (2 marks)
c) Suggest a *classification algorithm* to overcome the disadvantage of NBC. (1 mark)
d) The test data of a car is given as follows. Use NBC and the train data to determine whether or not the car is stolen. (6 marks)
Test data: *Colour = "Red", Type="SUV", Origin="Domestic"*

Train data:

Car Plate	Colour	Type	Origin	Stolen?
AAA1234	Red	Sports	Domestic	Y
BBB1235	Red	Sports	Domestic	N
CCC1236	Red	Sports	Domestic	Y
DDD1237	Yellow	Sports	Domestic	N
EEE1238	Yellow	Sports	Imported	Y
FFF1239	Yellow	SUV	Imported	N
HHH1240	Yellow	SUV	Imported	Y
III1241	Yellow	SUV	Domestic	N
JJJ1242	Red	SUV	Imported	N
KKK1243	Red	Sports	Imported	Y

Question 4

Five points are given as follows with x and y representing their locations.

A(1,8), B(7,2), C(2,1), D(8,4), E(7,0)

a) A and E are randomly selected centroids for Cluster1 and Cluster2, respectively. Use *Euclidean* as distance function, determine which cluster point C should belong to. (3 marks)
b) After the first round execution of k-means, assume that B,C, D and E belong to Cluster2. Draw a *scatter plot* for these five points. On the plot, group these points into two clusters manually. (2+1 marks)
c) What is the *new centroid* for Cluster2? (2 marks)
d) Can k-means handle outliers efficiently? Why or why not? (2 marks)

Continued...

Formulae:

Standard deviation, $\sigma = \sqrt{\frac{\sum_{i=1}^{n} x^2}{N} - \mu^2}$

Z-score normalization $v` = \frac{v-\mu}{\sigma}$

Covariance $Cov(A, B) = E(A \cdot B) - \bar{A}\,\bar{B}$

Correlation analysis $r_{A,B} = \frac{Cov(A,B)}{(n-1)\sigma_A a_B}$

NBC, $P(C_i|X) = P(X|C_i)P(C_i)$

Euclidean distance, $d(X, Y) = \sqrt{|x_2 - x_1|^2 + |y_2 - y_1|^2}$

End of Page